Amharic-English Information Retrieval with Pseudo Relevance Feedback

Atelach Alemu Argaw
Department of Computer and System Sciences, Stockholm University/KTH
atelach@dsv.su.se

Abstract

We describe cross language retrieval experiments using Amharic queries and English language document collection from our participation in the bilingual ad hoc track at the CLEF 2007. Two monolingual and eight bilingual runs were submitted. The bilingual experiments designed varied in terms of usage of long and short queries, presence of pseudo relevance feedback (PRF), and three approaches (maximal expansion, first-translation-given, manual) for word sense disambiguation. We used an Amharic-English machine readable dictionary (MRD) and an online Amharic-English dictionary in order to do the lookup translation of query terms. In utilizing both resources, matching query term bigrams were always given precedence over unigrams. Out of dictionary Amharic query terms were taken to be possible named entities in the language, and further filtering was attained through restricted fuzzy matching based on edit distance. The fuzzy matching was performed for each of these terms against automatically extracted English proper names. The Lemur toolkit for language modeling and information retrieval was used for indexing and retrieval. Although the experiments are too limited to draw conclusions from, the obtained results indicate that longer queries tend to perform similar to short ones, PRF improves performance considerably, and that queries tend to fare better when we use the first translation given in the MRD rather than using maximal expansion of terms by taking all the translations given in the MRD.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries

General Terms

Measurement, Performance, Experimentation

Keywords

Cross Language Information Retrieval, Amharic, Query Analysis

1 Introduction

Amharic is a Semitic language that is spoken in Ethiopia by an approximated 20-30 million people. It is a syllabic language, and uses a script which originated from the Ge'ez alphabet (the liturgical language of the Ethiopian Orthodox Church). The language has 33 basic characters with each

having 7 forms for each consonant-vowel combination, and extra characters that are consonant-vowel-vowel combinations for some of the basic consonants and vowels. It also has a unique set of punctuation marks and digits. Unlike other related Semitic languages such as Arabic, Hebrew or Syrian, Amharic is written from left to right. Amharic alphabets are one of a kind and unique to Ethiopia.

Manuscripts in Amharic are known from the 14th century and the language has been used as a general medium for literature, journalism, education, national business and cross-communication. A wide variety of literature including religious writings, fiction, poetry, plays, and magazines are available in the language.

Amharic has a complex but fairly structured morphological properties. To give some highlights: Amhaic has a rich verb morphology which is based on triconsonantal roots with vowel variants describing modifications to, or supplementary detail and variants of the root form. A significantly large part of the vocabulary consists of verbs, which exhibit different morphosyntactic properties based on the arrangement of the consonant-vowel patterns. Amharic nouns can be inflected for gender, number, definiteness, and case, although gender is usually neutral. Adjectives behave in the same way as nouns, taking similar inflections, while prepositions are mostly bound morphemes prefixed to nouns. The definite article in Amharic is also a bound morpheme, and attaches to the end of a noun.

The Amharic topic set for CLEF 2007 was constructed by manually translating the English topics by translators who are not involved in the retrieval tasks. The Amharic topic set which was written using the Ethiopic script (*fidel*), the writing system for Amharic, was then transliterated to an ASCII representation.

The two monolingual English retrieval experiments were conducted for comparison purposes. One used short queries containing the title and description fields of the English topic sets, while the other used long queries that contained title, description, and narrative fields of the topics. Two of the eight bilingual retrieval experiments conducted used short Amharic queries while the remaining six used long ones. The experiments also differed from one another in terms of the WSD method used and the use of pseudo relevance feedback in order to expand query terms. For indexing and retrieval, the Lemur toolkit for language modeling and information retrieval[1] was used.

The paper is organized as follows; Section 1 gives an introduction of the language under consideration and the overall experimental setup. Section 2 deals with the different steps taken in the query analysis. Section 3 describes how out of dictionary terms were handled, followed by approaches for word sense disambiguation in section 4. Section 5 discusses pseudo relevance feedback, and section 6 presents details about the designed experiments and the obtained results. These results are discussed and future directives are given in the last section.

2 Query Analysis

The query analysis starts with transliterating the Amharic script into an ASCII format. Stemming of the terms was then performed in order to handle morphological variations and insure that we find matches with the citation forms in the dictionaries for as many of the query terms as possible. Term bigrams were then looked up in the dictionaries and stop words were removed from the remaining Amharic query words based on corpus statistics. Remaining unigrams were then looked up in the dictionaries, giving a list of translation equivalents in English and unmatched terms to be considered for fuzzy matching. English stop words were also removed after the lookup translation using a publicly available stop words list for English. Each of these processes are described in more detail in this section.

[1] http://www.lemurproject.org/

2.1 Transliteration

The Amharic queries were written in *fidel*. For ease of use and compatibility purposes, the text was transliterated to an ASCII representation using SERA[2]. The transliteration was done using a file conversion utility called g2[3] which is available in the LibEth[4] package.

2.2 Stemming

We used an in-house developed software for stemming the Amharic query terms. The stemmer is designed to reduce morphological variants of words to their citation forms as found in the MRD. It finds all possible segmentations of a given word according to inflectional morphological rules of the language. Derivational variants are not handled since they tend to have separate entries in dictionaries. The most likely segmentation for the words is then selected based on occurrence statistics in a list of citation forms compiled from three dictionaries (Amharic-English, Amharic-Amharic, Amahric-French) and a 3.1 million words Amharic news corpus. The process is to strip off allowed prefixes and suffixes and look up the remaining stem (or alternatively, some morphologically motivated variants of it) in the list of citation forms to verify that it is a possible segmentation. Stem length is also taken into consideration when further disambiguation is needed. In the cases where stems cannot be verified using the dictionary lists, frequency of occurrence in the news corpus is used to decide which segmentation to pick. See [2] for a detailed information about the stemming process.

Bigrams are handled in the same manner, but the segmentation works in such a way that prefixes are removed from the first word and suffixes from the second one only. Compound words in Amharic are usually written as two words, but there is no inflection present as the suffix of the first word and prefix of the second word in the bigram.

2.3 Lookup Translation

The query translation was done through term-lookup in an Amharic-English MRD [1] and an online dictionary[5]. The machine readable dictionary contains 15,000 Amharic words and their corresponding English translations while the online dictionary contains about 18,000 entries. The lookup is done in such a way that the MRD translations are given precedence over the online dictionary translations, which are entered by users of the system and come with no guarantee as to their quality or correctness. Although this is the case, it should be noted that we have found the online dictionary to be quite useful and with good standard translations.

The lookup translation is done in the order that bigrams were looked up in the MRD, followed by bigram lookup in the online dictionary for those bigrams where no match is found in the MRD. In the next step, stop words were removed from the remaining terms (see following section) and unigrams were looked up in the MRD followed by a lookup of unigrams in the online dictionary if no match is found in the MRD. In all cases, when a match is found, all senses and synonyms of the term translations as given in the dictionaries were taken.

2.4 Stop Word Removal

Non content bearing words (stop words) were removed both before and after the lookup translation. First, all bigrams were extracted and looked up. The stop words were removed after excluding the bigrams for which matches were found in the dictionaries. This was done to ensure that we are not missing any possible bigrams due to removed stop words that are part of a meaningful unit. Before translation, Amharic stop words were removed based on global and local occurrence statistics. Each word's occurrence frequency was collected from the 3.1 million words news text,

[2]SERA stands for System for Ethiopic Representation in ASCII, http://www.abyssiniacybergateway.net/fidel/sera-faq.html

[3]g2 was made available to us through Daniel Yacob of the Ge'ez Frontier Foundation (http://www.ethiopic.org/)

[4]LibEth is a library for Ethiopic text processing written in ANSI C http://libeth.sourceforge.net/

[5]http://www.amharicdictionary.com/

and words with frequencies above 5,000 were considered to be stop words and are removed from the terms list. The remaining words were further checked by looking at their occurrence frequency in the 50 queries used. If they occur more than 15 times, they were also removed. The later stop word removal handled non content bearing words that are present in queries such as 'find', 'document', 'relevant' etc, which tend to have low occurrence frequencies in the news corpus.

English stop words were removed after the lookup translation. We used an English stop words list that comes with the Lemur toolkit, which is also used during the indexing of the English document collection.

3 Fuzzy Matching for Out of Dictionary Terms

Amharic query terms that are most likely to be named entities were selected automatically for fuzzy matching. Such words are query words that are not removed as stop words but for which no bigram or unigram match is found in both dictionaries. The unsegmented word form was retained for fuzzy matching and very commonly occurring noun prefixes and suffixes are stripped off. Prefixes such as 'be','ye','ke', and 'le', were removed when they are attached preceding a word and suffixes 'oc', 'oc-n', 'oc-na', 'oc-n-na' when they appear as the word endings.

Automatically extracting named entities for Amharic is difficult compared to that of English. Proper names in Amharic scripts are not capitalized. The absence of syntactic analyzer, a list of named entities, or a manually tagged text also makes it difficult (or time consuming if the resources are to be constructed from scratch) to train or base automatic named entity extraction with. Hence, in these experiments we opted for making use of features in the target language. We implemented a very simple and straight forward proper name extraction utility for English. We made use of the English document collection to extract these proper names, which included names of persons, organizations, places, awards, historical events, etc that begin with capital letters in the English document collection. Proper names that appear at the beginning of a sentence were not extracted since the capitalization at the beginning of a sentence is not always indicative of a proper name. We ensure that there isn't much 'noise' by discarding all sentence beginning words and although we might be missing out on some proper names, our assumption is that, if they occur ones, they tend to reappear elsewhere in the same text.

The extracted English proper names were then used for the subsequent process of fuzzy matching. An edit distance based fuzzy matching was done for the Amharic out of dictionary query terms that were selected to be possible named entities. Restricting the fuzzy matching to the extracted English proper names only rather than the entire document collection is believed to increase precision of the matches, while it lowers recall. We further restricted the fuzzy matching to contain terms with very high similarity levels only by setting the maximum allowed edit distance to be 2. Amharic terms for which no fuzzy match is found were removed while the shortest edit distance or preferred match is taken to be the English equivalent proper name for those words for which matches are found through the fuzzy matching. The preferred match is the match for which a predefined character in the Amharic word as given by the transliteration system [6] corresponds to a specific one in English. For example the Amharic transliteration 'marc' would have a 0 edit distance with the English proper name 'Marc' since we use lower cases for the fuzzy matching. But the English word 'March' which has an edit distance of 1 with the Amharic word 'marc' would be preferred since the Amharic 'c' in SERA corresponds to the sound 'ch' in English.

4 Word Sense Disambiguation

During the lookup translation using both dictionaries, all the senses given in the dictionaries for each term's translation were taken. In such a case, where there is no sense disambiguation and every term is taken as a keyword, we consider the queries to be 'maximally expanded' with all available senses and synonyms. The sense disambiguation in this case is left to be implicitly handled by the retrieval process. Some of the experiments discussed in the section below used the

'maximally expanded' set of translated keywords. Another set of experiments made use of only the first translation given in the dictionaries. Such an approach is an attempt to a very simplified and 'blind' word sense disambiguation, with the assumption that the most common sense of a word tends to be first one on the list of possible translations given in dictionaries. A manual sense disambiguation was also done for comparative purposes, to determine the effect of optimal WSD in the case of MRD based CLIR. Two of the reported experiments made use of the manually disambiguated set of keywords .

5 Pseudo Relevance Feedback

Pseudo Relevance Feedback (PRF) is a method of automatic local analysis where retrieval performance is expected to improve through query expansion by adding terms from top ranking documents. An initial retrieval is conducted returning a set of documents. The top n retrieved documents from this set are then assumed to be the most relevant documents, and the query is reformulated by expanding it using words that are found to be of importance (high weights) in these documents. PRF has shown improved IR performance, but it should also be noted that there is a risk of query drift in applying PRF[4]. Four of the experiments used PRF by including the highest weight 20 terms from the top ranking 20 documents, with a positive coefficient[6] of 0.5.

6 Experiments and Results

For indexing and retrieval, the Lemur toolkit for language modeling and information retrieval was used. The selection of this tool was primarily to try out language modeling approaches in Amharic-English cross language IR. We found that it was difficult to find optimal settings for the required smoothing parameters in the time frame allocated for this project, hence we reverted to the vector space models. Stop words were removed, and the Porter stemmer was used for stemming during indexing. Both features are available through the toolkit.

In information retrieval overall performance is affected by a number of factors, implicitly and explicitly. To try and determine the effect of all factors and tune parameters universally is a very complicated task. In attempting to design a reasonably well tuned retrieval system for Amharic queries and English document collections, our efforts lie in optimizing available resources, using language specific heuristics, and performing univariate sensitivity tests aimed at optimizing a specific single parameter while keeping the others fixed at reasonable values. In these experiments, we tried to see the effects of short queries vs. long queries, the use of PRF, and the effect of taking the first translation given versus maximally expanding query terms with all translations given in dictionaries.

What we refer to as long queries consisted of the title, description, and narrative fields of the topics, while short queries consisted of title and description fields. In the long queries, we filtered out the irrelevant info from the narrative fields, using cue words for Amharic. Amharic has the property that the last word in any sentence is always a verb, and Amharic verbs have negation markers as bound morphemes that attach themselves as prefixes onto the verbs. This property of Amharic has helped us in automatically determining whether or not a sentence in the narrative field of the topics is relevant to the query. Some of the sentences in the narrative fields of the topics describe what shouldn't be included or is not relevant for the query at hand. If we include all the sentences in the narrative fields, such information could possibly hurt performance rather than boost it. Therefore we looked at the last word in each Amharic sentence in the narrative field and removed those that have ending verbs marked for negation. Examples of such words used include 'ayfelegum', 'aydelum', 'aynoracewm' representing negations of words like 'needed', 'necessary', etc.

[6]The coefficient for positive terms in (positive) Rocchio feedback.

Table 1: Recall-Precision tables for the eight bilingual runs

Recall	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8
0.00	17.89	23.54	22.15	26.74	21.79	26.08	25.94	33.01
0.10	15.61	20.06	14.33	19.32	17.24	22.18	20.14	23.40
0.20	13.07	16.18	12.49	16.45	13.90	18.57	14.47	18.54
0.30	11.08	13.85	10.00	13.97	11.24	15.43	12.62	16.76
0.40	9.33	11.85	8.32	11.79	9.28	13.38	10.67	14.48
0.50	7.68	10.68	7.30	10.59	7.80	11.66	9.74	13.42
0.60	6.01	7.83	6.26	9.08	6.44	8.66	7.99	9.81
0.70	5.02	6.60	5.61	8.15	5.05	7.94	6.65	8.51
0.80	3.58	5.59	4.40	6.54	4.16	6.67	5.20	7.19
0.90	2.54	4.35	3.10	4.78	3.15	4.97	3.42	5.26
1.00	2.16	2.74	2.30	3.03	2.65	3.56	2.76	3.38

6.1 Designed Experiments

The experiments designed are:

- Run 1: Maximally expanded long queries (Title + Description + Filtered Narrative) were used.
- Run 2: Maximally expanded long queries, supplemented by PRF.
- Run 3: Maximally expanded short queries (Title + Description) were used.
- Run 4: Maximally expanded short queries, supplemented by PRF.
- Run 5: Long queries with word sense disambiguation using the first-translation-given approach.
- Run 6: Long queries with word sense disambiguation using the first-translation-given approach, supplemented by PRF.
- Run 7: Long queries with manual word sense disambiguation.
- Run 8: Long queries with manual word sense disambiguation, supplemented by PRF.

6.2 Results

The results obtained for the experiments discussed above are given in tables 1, 2 and 3. Table 1 presents precision values at different recall levels for the eight bilingual runs. Table 2 summarizes the results for these runs by presenting the number of relevant documents, the retrieved relevant documents, the non-interpolated average precision as well as the precision after R (where R is the number of relevant documents for each query) documents retrieved (R-Precision). Table 3 gives a summary similar to that of Table 2 for the monolingual English runs that were performed for comparison purposes.

7 Discussion and Future Directives

As can be seen in the results presented above, the best performance obtained was from the manually disambiguated word senses, followed by the first-translation-given approach, while the maximal expansion comes last. Long queries, that are believed to carry more information since they have a lot more keywords, were expected to perform much better than the shorter queries, but the results show that they have comparable performance. The automatic filtering of sentences

Table 2: Summary of results for the bilingual runs

	Relevant-tot	*Relevant-retrieved*	*Avg Precision*	*R-Precision*
`Run 1`	2247	880	7.77	8.78
`Run 2`	2247	951	10.5	10.44
`Run 3`	2247	873	7.71	8.21
`Run 4`	2247	943	10.97	10.57
`Run 5`	2247	868	8.29	10.17
`Run 6`	2247	1030	11.75	12.87
`Run 7`	2247	1002	9.75	10.85
`Run 8`	2247	1104	12.92	13.3

Table 3: Summary of results for the monolingual English runs

	Relevant-tot	*Relevant-retrieved*	*Avg Precision*	*R-Precision*
`Run 0`	2247	1399	22.84	24.47
`Run L`	2247	1435	24.05	25.49

in the narrative fields for long queries performed very well, removing all non-relevant sentences. Although that is the case, most of the additional information gained by using the long queries was a repetition to what is already been available in the short ones, except for a few additions. Using the narrative field also boosts negative impact through wrong segmentation and lookup. In depth analysis of a larger set of queries might shade some light into the positive and negative impact, although we believe that it still would be hard to draw conclusions from.

The use of PRF in all cases showed a substantial increase in performance. Given that the original retrieval precision is very low, it is very encouraging to see that PRF helps in boosting performance even in such cases. We plan to further pursue using PRF, and tuning parameters pertaining to PRF.

Amharic terms that have no match in the dictionaries were assumed to be named entities. Since the amount of entries in the two dictionaries utilized is 15,000 and 18,000 with possible overlaps, all out of dictionary entries would not possibly be named entities. In order to handle this issue, the fuzzy matching is restricted to English proper names only and a very high similarity requirement was set for the fuzzy matching supplemented by language specific heuristics. We intend to investigate this further by looking at ways of bootstrapping a named entity recognizer for Amharic, especially following the approaches discussed for Arabic by [5], as well as using a more sophisticated named entity recognizer for English to extract as many named entities as possible, rather than restrict it to proper names only.

The fact that manual WSD gave the best results and that blindly picking the first translation given has better performance than maximal MRD expansion of query terms motivates us to put more effort in investigating approaches to automatic WSD. Given the resource limitations, the best approach is most likely to use target language document collection and contextual collocation measures for sense disambiguation. We intend to investigate further approaches presented in [3] as well as experiment with a few more collocation measures.

Stemming plays a crucial role in MRD based CLIR since whether we would find the correct match in the dictionary depends on how well the stemmer does. We will pursue further attempts made so far to optimize the performance of the stemmer.

Although the results obtained are indicative of the facts presented above, the experiments are too limited to draw any conclusions. Large scale experiments using a larger set of queries and data set including those from previous years of CLEF ad hoc tasks will be designed in order to give the results more statistical significance. The relatively low precision levels are also issues we

plan to investigate further by taking a closer look at the indexing and retrieval experiments.

References

[1] Amsalu Aklilu. *Amharic English Dictionary.* Mega Publishing Enterprise, Ethiopia, 1981.

[2] Atelach Alemu Argaw and Lars Asker. An amharic stemmer : Reducing words to their citation forms. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, pages 104–110, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[3] Atelach Alemu Argaw, Lars Asker, Rickard Cster, Jussi Karlgren, and Magnus Sahlgren. Dictionary-based amharic-french information retrieval. In Carol Peters, Fredric C. Gey, Julio Gonzalo, Henning Mller, Gareth J. F. Jones, Michael Kluck, Bernardo Magnini, and Maarten de Rijke, editors, *CLEF*, volume 4022 of *Lecture Notes in Computer Science*, pages 83–92. Springer, 2005.

[4] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval.* Cambridge University Press, 2008.

[5] Khaled Shaalan and Hafsa Raza. Person name entity recognition for arabic. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, pages 17–24, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[6] D. Yacob. System for ethiopic representation in ascii (sera). http://www.abyssiniacybergateway.net/fidel/, 1996.